Kontingencia táblák. Khi-négyzet teszt

1. **Függetlenségvizsgálat**
2. **Illeszkedésvizsgálat**
3. **Homogenitásvizsgálat**

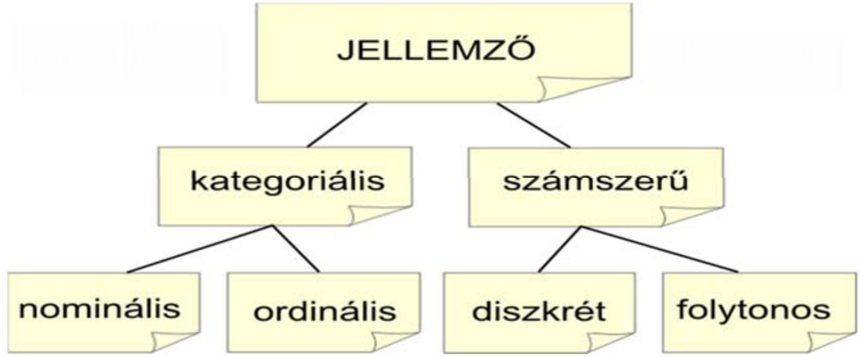
Példa 1

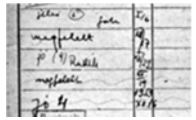
	szemü-veges	nem sz.	össze-sen
nő	28	75	103
fér-fi	48	49	97
	76	124	200

?

KAD 2015.11.19

Ismétlés: változók, mérési skálák típusai





jeles (5)
jó (4)
közepes (3)
elégséges (2)
elégtelen (1)

---

1. Függetlenségvizsgálat

Kapcsolatvizsgálat kategorikus változók között. Khi-négyzet teszt

gyakorisági táblázat (kontingencia táblázat): két változó közös gyakoriságának táblázatos ábrázolása
X (pl. nem) és *Y* (szemüvegesség)

	szemü-veges	nem sz.	össze-sen
nő	a=28	b=75	103
fér-fi	c=48	d=49	97
	76	124	200

kérdés: különbözik-e egy rögzített tulajdonság gyakorisága a két csoportban?

A nullhipotézis felállítása

H_0: nem és szemüvegesség egymástól függetlenek
(nincs különbség a csoportokban)

$$\frac{a}{b} = \frac{c}{d}, \quad \text{vagy} \quad \frac{a}{c} = \frac{b}{d}$$

mekkora lenne a **várt gyakoriság** (expected frequency) a bal felső (*a*) cellában, ha a nullhipotézis igaz?

a nők száma:
$a + b = 103$
a szemüveges személyek száma:
$a + c = 76$

a nők aránya a mintában:
$p(\text{nő}) = (a + b)/n = 103/200$
a szemüvegesek aránya a mintában :
$p(\text{szemüveges}) = (a + c)/n = 76/200$

	szemü-veges	nem sz.	össze-sen
nő	a=28	b=75	103
fér-fi	c=48	d=49	97
	76	124	200

a megfigyelt (observed) gyakoriságok táblázata

Várt gyakoriságok. feltevés: H_0 igaz ⇨ a nem és a szemüvegesség független tulajdonságok

várt gyakoriság a bal felső cellában : $\frac{a+b}{n}\cdot\frac{a+c}{n}\cdot n=\frac{(a+b)\cdot(a+c)}{n}$

várt gyakoriság a jobb felső cellában : $\frac{a+b}{n}\cdot\frac{b+d}{n}\cdot n=\frac{(a+b)\cdot(b+d)}{n}$

várt gyakoriság a bal alsó cellában : $\frac{c+d}{n}\cdot\frac{a+c}{n}\cdot n=\frac{(c+d)\cdot(a+c)}{n}$

várt gyakoriság a jobb alsó cellában : $\frac{c+d}{n}\cdot\frac{b+d}{n}\cdot n=\frac{(c+d)\cdot(b+d)}{n}$

	sz.	nem	össz.
n	a=28	b=75	103
f	c=48	d=49	97
	76	124	200

megfigyelt (observed) kontingencia táblázat

	sz.	nem	össz.
n	103*76/200	103*124/200	103
f	97*76/200	97*124/200	97
	76	124	200

várt (expected) kontingencia táblázat

A várt gyakoriságok a megfigyelt gyakoriságokból

	sz.	nem	össz.
n	a=28	b=75	103
f	c=48	d=49	97
	76	124	200

megfigyelt (observed) kontingencia táblázat

	sz.	nem	össz.
n	103*76/200	103*124/200	103
f	97*76/200	97*124/200	97
	76	124	200

várt (expected) kontingencia táblázat

$$(\text{várt gyakoriság})=\frac{(\text{oszlopösszeg})\cdot(\text{sorösszeg})}{(\text{a minta elemszáma})}$$

Próbastatisztika

Ha a nullhipotézis igaz:
A megfigyelt és a várt gyakoriságokat tartalmazó kontingencia táblázatok megfelelő celláiban levő értékek nagyjából egyformák. A következő próbastatisztika (súlyozott négyzetes közép) **khi-négyzet eloszlású**:

$$\boxed{\chi^2=\sum_i\frac{(O_i-E_i)^2}{E_i},}$$

ahol
O_i a megfigyelt (observed)
E_i a(z el)várt gyariságok az i-dik cellában.
Szabadsági fok: (sorok száma –1)*(oszlopok száma –1),
ill. egydimenziós esetben: (cellák száma –1)

pl. 2*2 (négymezős-) táblázat: 1

A teszt végrehajthatóságának feltételei

n (a minta elemszáma) elegendően nagy:

a *várt gyakoriságokat* tartalmazó kontingencia táblázatban minden cellatartalomnak 1-nél nagyobbnak kell lenni

a *várt gyakoriságokat* tartalmazó kontingencia táblázatban azoknak a celláknak a száma, amelyekben a cellatartalom 1 és 5 közötti csak a cellák 20%-a lehet

(pl. négymezős táblázat: minden cellában a cellatartalomnak 5-nél nagyobbnak kell lenni)

Speciális eset: négymezős táblázat (gyakorlati jegyzet 2.b.29)

	a vizsgált tulajdonság		összesen
	megvan	nincs meg	
A csoport	***a***	***b***	*a*+*b*
B csoport	***c***	***d***	*c*+*d*
összesen	*a*+*c*	*b*+*d*	*n*

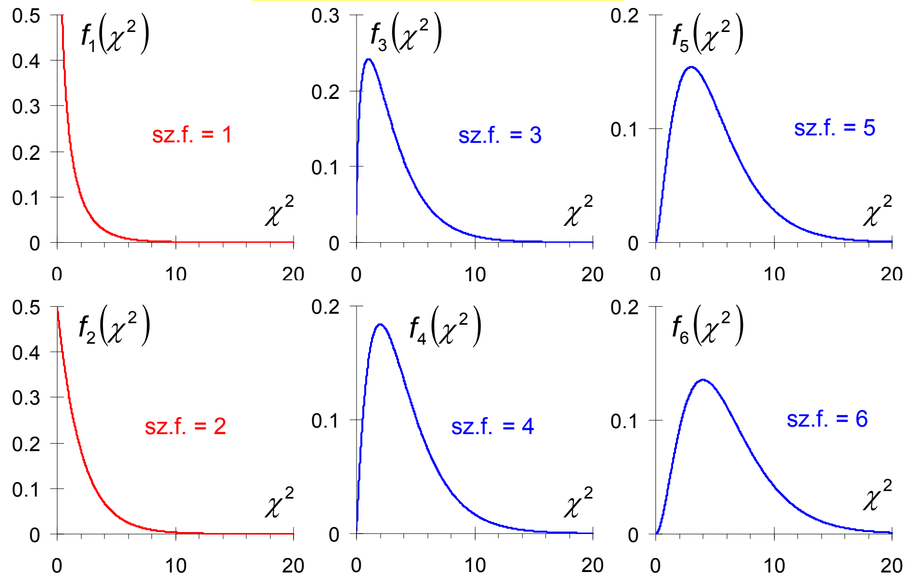
$$\chi^2_{\mathrm{M}} = \frac{n\cdot(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

a végrehajthatóság feltétele:

a két legkisebb részösszeg szorzata legyen nagyobb, mint 5*n*

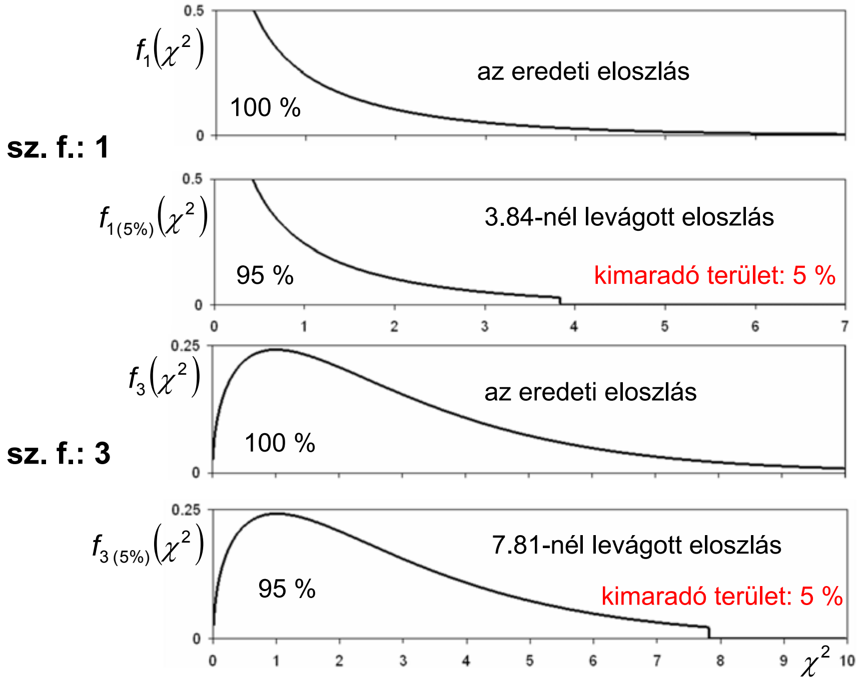
$$\frac{a}{b}=\frac{c}{d} \quad \Leftrightarrow \quad ad=bc$$

Khi-négyzet eloszlások



módusz = 0, ha sz.f.=1 vagy 2

módusz = (sz.f.-2), ha sz.f. >2



szabadsági fok	*p* (valószínűség) 0,99	0,975	0,95	0,05	0,025	0,01	0,001
1	0,0000157	0,0000982	0,000393	3,84	5,02	6,63	10,83
2	0,0201	0,0506	0,103	5,99	7,88	9,21	13,82
3	0,115	0,216	0,352	7,81	9,35	11,34	16,27
4	0,297	0,484	0,711	9,49	11,14	13,28	18,47
5	0,554	0,831	1,15	11,07	12,83	15,09	20,51
6	0,872	1,24	1,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	15,51	17,53	20,09	26,13
9	2,09	2,70	3,33	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	18,31	20,48	23,21	29,59
11	3,05	3,61	4,57	19,68	21,92	24,72	31,26
12	3,57	4,40	5,23	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	23,68	26,12	29,14	36,12

Példa 1

A teszt alkalmazhatóságának feltétele:
a két legkisebb részösszeg szorzata legyen nagyobb, mint 5n

	szemü-veges	nem sz.	össze-sen
nő	a=28	b=75	103
fér-fi	c=48	d=49	97
	76	124	200

76*97 = 7372 > 5*200 =1000

a khi-négyzet teszt használható

$$\chi^2_M = \frac{200 \cdot (28 \cdot 49 - 48 \cdot 75)^2}{76 \cdot 124 \cdot 103 \cdot 97} = 10.54$$

10.54 > χ^2_{krit} =3,84 H_0 hamis

van kapcsolat a nem és a szemüvegesség (szemüvegviselési hajlandóság!) között

szabad sági fok	p (valószínűség) 0,99	0,975	0,95	0,05	0,025	0,01	0,001
1	0,0000157	0,0000982	0,000393	3,84	5,02	6,63	10,83

$$\chi^2_M = \frac{200 \cdot (28 \cdot 49 - 48 \cdot 75)^2}{76 \cdot 124 \cdot 103 \cdot 97} = 10.54$$

10.54 > χ^2_{krit} =3.84 H_0 hamis

10.54 > χ^2_{krit} =6.63 H_0 hamis

elvetjük a nullhipotézist, szignifikancia szint: <0.01

	A	B	C	D
1	megfigyelt gyakoriságok			
2		szemüveges	nem sz.	
3	nő	28	75	=SUM(B3:C3)
4	férfi	48	49	=SUM(B4:C4)
5		=SUM(B3:B4)	=SUM(C3:C4)	=SUM(D3:D4)
6	várt gyakoriságok			
7		szemüveges	nem sz.	
8	nő	=$D3*B$5/$D$5	=$D3*C$5/$D$5	=SUM(B8:C8)
9	férfi	=$D4*B$5/$D$5	=$D4*C$5/$D$5	=SUM(B9:C9)
10		=SUM(B8:B9)	=SUM(C8:C9)	=SUM(D8:D9)
11			szignifikanciaszint:	=CHITEST(B3:C4,B8:C9)
12			khi-négyzet érték:	=CHIINV(D11,1)

	A	B	C	D
1	megfigyelt gyakoriságok			
2		szemüveges	nem sz.	
3	nő	28	75	103
4	férfi	48	49	97
5		76	124	200
6	várt gyakoriságok			
7		szemüveges	nem sz.	
8	nő	39.14	63.86	103
9	férfi	36.86	60.14	97
10		76	124	200
11		szignifikanciaszint:		0.0012
12		khi-négyzet érték:		10.544

számolás Excel-lel

angol		*magyar*
SUM	=	SZUM
CHISQ.TEST	=	KHI.PRÓBA
CHIDIST.RT	=	KHI.ELOSZLÁS
CHISQ.INV.RT	=	KHI.INVERZ.JOBB

példa 2

	sz.	nem sz.	ösz-szes
nő	1	3	4
férfi	5	3	8
	6	6	12

?

4*6 = 24 < 5*12 =60

a khi-négyzet teszt nem használható

(helyette: Fisher egzakt teszt)

a minta elemszámának növelése

	sz.	nem	össz.
nő	1	3	4
férfi	5	3	8
	6	6	12

12 → 200

	sz.	nem	össz.
nő	28	75	103
férfi	48	49	97
	76	124	200

nők: $\frac{n_{sz}}{n_{nem}} = \frac{1}{3} = 0.33$ $\qquad$ $\frac{n_{sz}}{n_{nem}} = \frac{28}{75} = 0.37$

férfiak: $\frac{n_{sz}}{n_{nem}} = \frac{5}{3} = 1.67$ $\qquad$ $\frac{n_{sz}}{n_{nem}} = \frac{48}{49} = 0.98$

sejtésünk van, de nem tudjuk igazolni

n növelésével (12 → 200): a sejtés igazolható lesz

Példa 3 (biofizika jegyzet 102. példa). Nem artériás típusú ischaemiás opticus neuropathia sikeres műtéti korrekciójáról jelent meg 1989-ben egy közlemény. Minthogy e betegségben korábban semmiféle hatásos kezelési módszer nem volt ismert, ezt a műtétet sok helyen alkalmazni kezdték. Rövidesen eredménytelen beavatkozásokról is megjelentek beszámolók, ezért számbavették 25 klinikai centrum 244 ilyen betegét, akik közül 119 főn elvégezték a műtétet, 125 betegen nem. A felmérés eredménye:

megfigyelt gyakoriságok

	műtött	nem m.	össz.
javult	39	53	92
változatlan	52	56	108
romlott	28	16	44
összes	119	125	244

várt gyakoriságok

	műtött	nem m.	össz.
javult	45	47	92
változatlan	53	55	108
romlott	21	23	44
összes	119	125	244

khi^2 = $(39–44.87)^2/44.87+(53–47.13)^2/47.13 +(52–52.67)^2/52.67+(56–55.33)^2/55.33 +(28–21.46)^2/21.46+(16–22.54)^2/22.54 = 5.407$

Mivel $5.407 < 5.991= \chi^2_{krit,\ sz.f.=2}$, ezért nem vethetjük el a nullhipotézist. Azaz a mintánk alapján nincs okunk feltételezni különbséget a két módszer (műtét ill. nem műtét) hatásossága között.

egydimenziós kontingencia táblázatokkal kapcsolatos kérdés:
a megfigyelt értékek illeszkednek-e egy feltételezett eloszláshoz?

2. Illeszkedésvizsgálat (goodness of fit). Khi-négyzet teszt

tiszta illeszkedésvizsgálat

(a gyakoriságokat ismert valószínűségekből kapott gyakoriságokkal hasonlítjuk össze)

- egyenletes eloszlásra történő i.v.
- egyéb ismert paraméterű eloszlásra történő i.v.

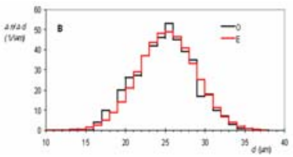
kockafeldobás eredménye

1	2	3	4	5	6
21	14	14	19	16	16

becsléses illeszkedésvizsgálat

(az eloszlás típusa alapján a megfigyelt gyakoriságokból becsüljük az eloszlás paramétereit)

- normalitás-vizsgálat
- egyéb becsült paraméteres i.v.



Egyenletes eloszlásra történő illeszkedésvizsgálat

A megfigyelt gyakoriságokat tartalmazó kontingencia táblázatot (bekeretezett rész, *O*) kibővítjük a várt gyakoriságokat tartalmazó segéd-kontingencia táblázattal (*E*). Feltételezzük, hogy a kocka nem cinkelt (H_0), ezért a 6 lehetséges esemény egyforma gyakoriságú: 100/6 = 16.7

a kockafeldobás eredménye

	1	2	3	4	5	6	össz.
O	21	14	14	19	16	16	100
E	16.7	16.7	16.7	16.7	16.7	16.7	100

khi^2= $(21–16.7)^2/16.7 +(14–16.7)^2/16.7 + (14–16.7)^2/16.7+ +(19–16.7)^2/16.7 + (16–16.7)^2/16.7 +(16–16.7)^2/16.7 = =2.36 < 11.07= \chi^2_{krit,\ sz.f.=5}$, a nullhipotézist megtartjuk.
A kocka nem cinkelt.

Normalitás-vizsgálat

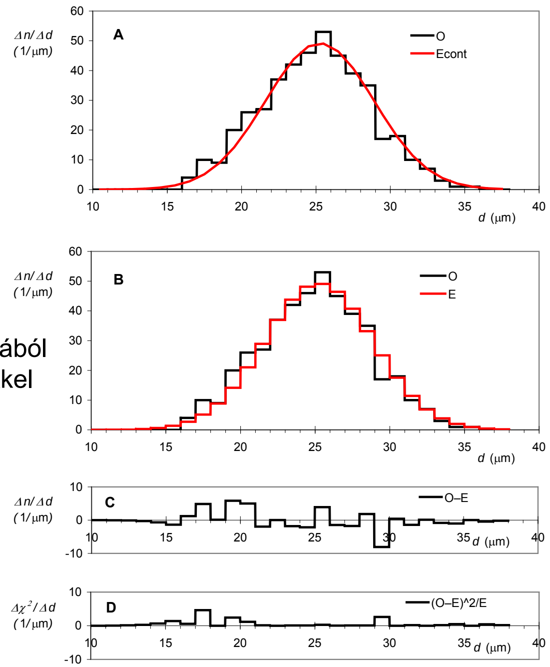
H_0: a béka vörösvérsejt hosszabbik átmérője normáleloszlású

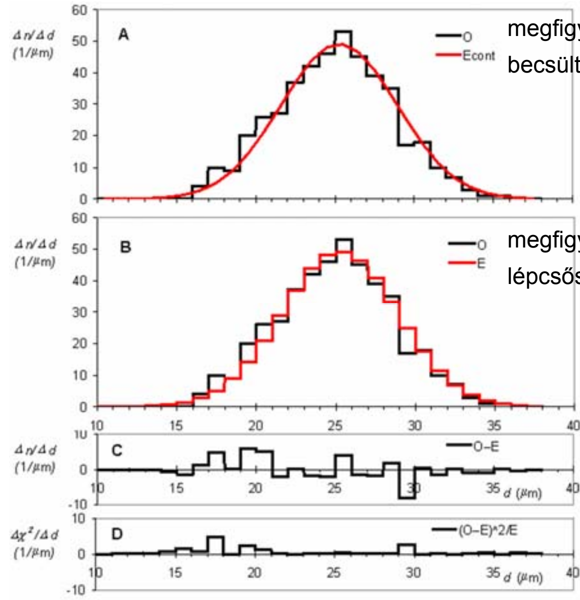
Az elméleti értékeket a mintából számolt tapasztalati értékekkel becsüljük.
Sz.f. = n-m-1, m: a becsült paraméterek száma (itt: 2)

...

p = 0.9 > 0.05

H_0: megtartjuk





megfigyelt gyakoriságok
becsült eloszlás sűrűségfüggvény

megfigyelt gyakoriságok
lépcsőssé tett becsült eloszlás ss. fv.

megfigyelt és várt függvények különbségei

khi-négyzet érték szemléletes ábrázolása (a görbe alatti terület)

3. **Homogenitásviszgálat,** (test for homogeneity)

H_0: a biofizika kollokviumjegyek eloszlása* a női hallgatók között ugyanolyan mint a férfi hallgatók között
(az eloszlások homogének)

kollokviumjegy, biofizika	nő	férfi	
5	22	12	34
4	26	31	57
3	27	38	65
2	23	25	48
1	14	13	24
	112	119	231

megfigyelt (observed) gyakoriságok

kollokviumjegy, biofizika	nő	férfi	
5	16.5	17.5	34
4	27.6	29.4	57
3	31.5	33.5	65
2	23.3	24.7	48
1	13.1	13.9	24
	112	119	231

várt (expected) gyakoriságok

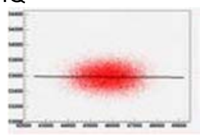
p = 0.27 > 0.05 H_0: megtartjuk

*adatok: 2009 őszi szemeszter

Függőségi viszonyok lehetőségei

függetlenség

IQ



testmagasság

függőség

sztochasztikus viszony

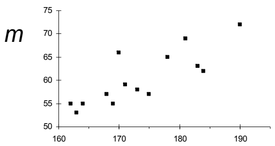
determinisztikus viszony

korrelációs

vegyes

asszociációs

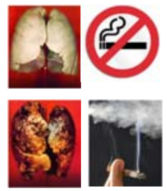
m



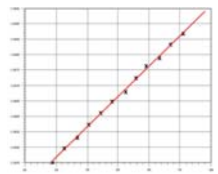
testmagasság

íz

színe-zettség



n



koncentráció

számszerű

ordinális

nominális

számszerű